Online Platform for Extracting, Managing, and Utilising Multilingual Terminology

Mārcis Pinnis, Tatiana Gornostay, Raivis Skadiņš, Andrejs Vasiļjevs

Tilde, Vienības gatve 75a, Riga, Latvia, LV-1004

{marcis.pinnis, raivis.skadins, tatiana.gornostay, andrejs}@tilde.lv

Abstract

In this demonstration paper we present an innovative platform "Terminology as a Service" (TaaS) for acquiring raw terminological data, and cleaning up, sharing, and reusing them, based on cloud computing. The platform serves, among other things, the needs of specialised lexicography. The proposed solution aims to fill the gap of collaborative terminology management and effective sharing of existing terminological data thus speeding up the development of specialised dictionaries. It also aims to build a bridge for the reuse of existing terminology between different groups of users, e.g. human users, such as lexicographers, translators, terminologists, and others, and machine users, such as computer-assisted translation tools, machine translation systems, third party terminology management solutions, and others.

Keywords: specialised lexicography, terminography, specialised dictionary, terminology service

1. Introduction

Lexicography, as the theory and practice of dictionary development, is one of the most labour-intensive human activities in the field of linguistics. The creation of a new dictionary from scratch and its delivery to an end user requires considerable resources in terms of time, man power, and finance. The main drawback of a conventional "paper" dictionary is its static and out-of-date content. For example, a particular paper terminological dictionary was already out-of-date by about 5–6 years when it was published and distributed (Shaikevich, 1983). In specialised lexicography, or terminography, it is even more critical since terminology is developing rapidly along with its subject field, or domain, and science in general.

Accurate handling of terminology is dramatically important in any professional language work—domain expertise, terminological analysis, documentation authoring and translation, professional (corporate and industry) communication, brand and product management, and other processes.

A paper terminological dictionary is somewhat a static fragment of a certain subject field in a certain language at a certain period of time.

To overcome the shortcomings of conventional lexicography, an electronic

punch-card machine was first used to create a prototype of a modern electronic dictionary by Roberto Busa in the XXth century. His first work was based on the automatic linguistic analysis (lemmatisation) of the works of Saint Thomas Aquinas. Roberto Busa compared the invention of an "electronic book" (instead of a printing book) to the introduction of a printing book by Gutenberg (instead of a manuscript) (Busa, 1961). Since that time automated lexicography has been developing rapidly.

Nowadays, with the evolution of information technologies, the Internet, and data (e.g., open data on the Web, free parallel and comparable corpora, and many other resources), the task of automated, or computational, specialised lexicography becomes a priority. Routine processes have been delegated to a computer. An electronic, or computer-based dictionary is easy to update and manage, and its main advantage is its flexible, dynamic, and extensible (e.g., in terms of new languages) character. Moreover, the new era of information technologies offers new ways of dictionary representation, e.g., on tablet, mobile, and other devices, and the usage patterns of a dictionary are changing with the course of time.

Lexicographers can have access to data and process them – analyse, tag, extract information etc. The integration of natural language processing tools have made it possible to grammatically and semantically analyse and tag a text and then to extract required pieces of information from the text. In the specialised lexicography, or terminography, developers can analyse and extract term candidates for further processing, e.g., automatic clean-up, sharing, and reusing in further processing and/or other applications (see section 3 and 4 below). Thus it has become possible to consider hundreds of thousands of terms specific to a certain domain in comparison with that time when only several thousands, usually no more than 2000, could be included in a conventional dictionary.

In this demonstration paper we present an innovative cloud-based platform "Terminology as a Service" (TaaS) for acquiring raw terminological data, cleaning it up, sharing and reusing terminological data cleaned up by users. The platform serves, among other thingss, the needs of specialised lexicography.

The proposed solution aims to fill the gap of collaborative terminology management and effective sharing of existing terminological data and thus speeding up the development of specialised dictionaries. It also aims to build a bridge for the reuse of existing terminology between different groups of users, e.g., human users, such as lexicographers, translators, and terminologists (human-oriented specialised dictionaries), as well as machine users, such as computer-assisted translation (CAT) tools, machine translation (MT) systems, third party terminology management solutions etc. (machine-oriented specialised dictionaries). The paper is structured as follows: section 2 provides a brief overview of the TaaS platform, section 3 describes the workflow for the creation of a bilingual terminological collection from user-provided documents, terminology sharing and reusing possibilities offered by

the platform are outlined in section 4, and available interfaces for machine users are briefly drafted in section 5. Finally, the paper is concluded and future work is outlined in section 6.

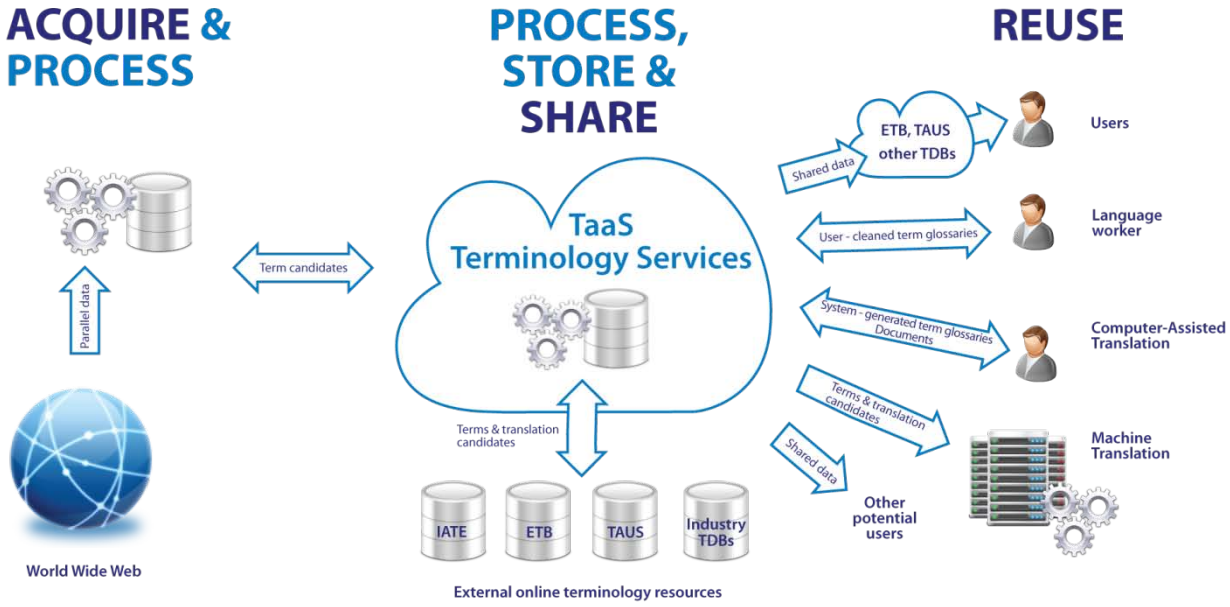

Figure 1: The concept of the innovative cloud-based platform for terminology services

2. TaaS in an Outline

The TaaS platform is being developed in an industry-research collaboration project within the EU Seventh Framework Programme for Research and Technological Development.

The motivation of the TaaS platform is to facilitate terminology work in practical translation scenarios by providing a number of online terminology services.

User surveys have shown that translators, editors, technical writers, and other language workers spend up to 30% of their working time on terminology research, looking for terms in multiple local and online sources, acquiring terminology, and organising proprietary terminology glossaries (Blancafort et al., 2010). In some cases, terminology research can consume more than 30% of overall working time, e.g. in the translation of technical specifications (Massion, 2007). A language worker usually needs immediate answers to terminology requests but due to time and cost constraints proper terminology search is often skipped. Resulting errors in term usage affect not only translation/localisation productivity and overall costs but also influence further stages of documentation life cycle, e.g., failures in product technical support, client request processing etc.

TaaS addresses these needs by establishing a cloud-based platform to provide online terminology services for key terminology tasks – term identification in user-uploaded documents, translation equivalent recognition in existing terminology resources for identified term candidates, terminological collection creation, acquisition of translation equivalent candidates from the Web (parallel and comparable Web resources), crowd-sourced clean-up of terminological data, sharing these data and reusing them in crucial usage scenarios, and thus becoming a part of the multifaceted global cloud-based service infrastructure.

The platform offers automated workflows and facilities for the following activities:

- Automatic identification of monolingual term candidates in user-uploaded documents using state-of-the-art linguistically and statistically motivated term extraction techniques.
- Automatic recognition of term translation equivalent candidates for terminological units identified in user-uploaded documents using the largest publicly available terminology databases, such as IATE[1] and EuroTermBank[2], as well as statistical terminological lists acquired from domain-specific comparable corpora and publicly available parallel corpora found on the Web.
- Collaborative terminology clean-up (creating, editing, deleting term entries) and monolingual and multilingual terminological collection creation.
- Sharing and reusing user-created and publicly available terminology (including monolingual and bilingual terminological collections) with the help of import/export application programming interfaces (APIs) for automated processes and easy-to-use graphical user interfaces for human users.

The TaaS platform provides also Web service-based interfaces for CAT tools and MT systems that allow terminology look-up in external terminology databases and user-created private and public terminological collections, specialised terminological collection retrieval and term translation candidate mark-up within translatable documents for CAT, MT, and other automated tasks. The conceptual design of the TaaS platform is depicted in Figure 1.

3. Workflow for the Creation of a Bilingual Terminological Collection

Translators, technical writers, terminologists, and other language workers, when working on domain-specific writing tasks (i.e., translation, documentation etc.), require in-domain terminological dictionaries (monolingual and multilingual) that

[1] http://iate.europa.eu/

[2] http://ww.eurotermbank.com

can aid them in their effort to produce content that simultaneously has correctly and consistently applied terminology. The TaaS platform provides a workflow for *Bilingual Terminology Collection Creation* that allows human users to create terminological collections (i.e., raw terminological dictionaries) semi-automatically from user uploaded documents and comparable and parallel corpora found on the Web.

3.1 Monolingual Term Extraction

The TaaS platform allows semi-automatic terminological collection creation from multiple key formats that have been identified in a prior target user survey described in Gornostay et al. (2013) as the most used by the community including the Open Document[3] formats, PDF and several parallel data exchange formats, e.g., TMX and XLIFF.

In the first step, plaintext is extracted from user-uploaded documents and terms are tagged in the documents with statistically and linguistically motivated term extraction methods following Pinnis et al. (2012) in three steps. At first, term candidates are acquired using part-of-speech pattern filtering. Then, terms are weighed using different statistical association measures; the weights are normalised with the help of the TF*IDF (Spärck Jones, 1972) measure using reference corpora statistics (i.e., an inverse document frequency list calculated on a broad domain corpus). The platform supports term tagging for all 23 official languages of the European Union and also for Russian and Croatian.

After term tagging, a monolingual terminological collection in the TBX[4] format is created. At first, all unique terms are extracted from the tagged documents and normalised (i.e., transformed from the term morpho-syntactic surface forms to the morpho-syntactic base forms). As term normalisation is a language dependent task, it is currently available for selected languages (including English, Hungarian, Latvian, Lithuanian, and other project languages). If normalisation is applied, monolingual terms are consolidated (i.e., different surface forms of the same term are grouped together as one term entry) using term normalised forms and the respective morpho-syntactic descriptions of the normalised forms. If normalisation is not applied, monolingual terms are consolidated using term lemma sequences and part-of-speech sequences.

When automated processes are completed, the user can perform terminology clean-up or execute term translation lookup in order to proceed to multilingual terminological collection creation.

[3] Open Document Format for Office Applications

[4] TBX is a terminology exchange format originally created by the Localization Industry Standards Association (LISA) and later standardised by ISO as international standard ISO 30042:2008.

3.2 Retrieval of Term Translation Equivalents

After extracting terms from user-provided documents, the TaaS platform creates a bilingual terminological collection by finding potential translation equivalents for each of the extracted terms. For this, TaaS queries several sources of terminological data looking for entries that match the search term, are in the same subject field, and have target language equivalents.

Four types of terminological data are queried:

- private (confidential) TaaS terminological collections of the particular user,
- terminological collections shared by other TaaS users,
- external terminology databases,
- and the TaaS Statistical Database, which contains translation equivalent candidates acquired from comparable and parallel corpora found on the Web.

Let us briefly describe the sources mentioned above. The TaaS platform provides facilities to store all terminological collections created by the user. The user can create a collection either by applying TaaS workflows on the user-provided documents or by importing his/her locally created dictionary into the TaaS platform. Common formats, such as TBX and CSV, are supported for importing user terminology.

By default, user terminology is private, i.e., accessible only to the user. The owner of the terminology can provide individual access rights to his/her terminology to other users within the working group.

We encourage users to share their terminology with other users by changing their status to *Shared* (public). By sharing their terms, users participate in a collaborative effort to increase the size and scope of publicly available terminology resources. Shared terminological collections are accessed and used by both TaaS users and by TaaS workflows.

TaaS also searches several well-established online terminology databases:

- EuroTermBank: an online multilingual terminology portal providing consolidated access to 2.6 million terms from 137 terminology resources in more than 30 languages (Vasiļjevs et al., 2008),
- IATE: an EU inter-institutional terminology database containing 1.4 million multilingual entries used in different EU legislative acts and other documents,
- TAUS Data Repository: a large collection of shared translation memories (TM) provided by members of TAUS (Translation Automation User Society).

It should be noted that TAUS translation memories consisting of sentences and text

segments with their translations cannot be considered as a terminological resource. But in some fields, such as information technology, TM include many terms and their translation originates from software interface and product documentation, and TM strings with exact match are retrieved by TaaS.

Querying terminology resources and TM is a relatively straightforward process. But as new terms in different areas are appearing very frequently and they have to be translated in many languages, even the best terminology databases include only a fraction of terms that are appearing in the daily workload of translators.

To assist in translating terms that do not have translation equivalents in terminology databases, TaaS provides means of finding possible translations in Web data. For this purpose TaaS collects parallel and comparable data from the Web, aligns it at sentence and word levels and extracts potential term candidates with their translation candidates. Comparable corpora consist of original source-target language document pairs on the same topic, thus not translations of each other.

For data collection and extraction, TaaS uses an updated version of the ACCURAT Toolkit (Pinnis et al., 2012). The ACCURAT Toolkit provides tools and workflows for acquisition and processing of comparable corpora in order to acquire multilingual parallel data (including terminological data). Parallel and comparable Web data are collected from multilingual news feeds, focused Web domains, and Wikipedia. This workflow runs periodically in the background and stores resulting terms in the TaaS Statistical Database.

3.3 Collaborative Terminology Clean-up

The progress in information technologies and their role in the modern specialised lexicography cannot be overestimated. However, a specialist is the one who decides whether a linguistic unit is a term or not. This is about the unithood and termhood of a term and is out of scope of this paper, although it is one of the important steps in a term life cycle. Professionals seek joint collaboration and exchange of terminological data, and the TaaS platform offers these functionalities. Several of the data clean-up functions that are provided by the TaaS platform are: deletion of term candidates from the terminological collections, editing of various data categories of term entries within the terminological collections, changing status of the term candidates etc.

4. Sharing and Reusing Terminology

The concept of sharing, unfortunately, is not present to a considerable extent in the current models of major terminology resources – instead of providing the opportunity for consumers to contribute, reuse, and share their data, major terminology resources (term banks and databases) typically keep to the traditional one-way communication of their high quality pre-selected content.

According to a recent survey, there is a need for collaborative solutions and sharing models – 60.5% of respondents (out of 1782 participants) are willing to share their terminology with colleagues (Gornostay et al., 2013).

The core objective of the TaaS platform is, however, to align the speed of terminology resource management with the speed at which multilingual documentation is created. In order to achieve this goal, the TaaS platform allows its users to take full control of their monolingual and multilingual terminological collections and lets them decide with whom to share their terminology, to whom to grant the rights of collaborative improvement of terminological content, and to whom to grant access rights to the user terminology.

The TaaS platform provides human users with simple terminological collection importing and exporting methods in TBX, CSV (comma-separated values), and TSV (tab-separated values) formats. When terminological collections are imported within the TaaS platform, they are immediately accessible to other third party systems that support the TaaS platform's API (provided that the user has access to the third party systems).

The user can also make his/her terminological collections completely public, thus sharing them with every user of the TaaS platform.

5. Interfaces for CAT Tools and MT Systems

Multilingual consolidated and harmonised terminology in the form of monolingual and multilingual terminological collections is already utilised as an important resource in the process of human translation. However, a dictionary user is not necessarily a human specialist but could be an automated system: a machine user. Therefore, the TaaS platform also offers access to multilingual terminological collections through a dedicated Web service API. Typical machine users that may benefit from the service are, for instance, CAT tools, MT systems, authoring and (multilingual) documentation and content management systems, terminology management systems, indexing systems, search engines, Web crawlers, information retrieval systems (including cross-lingual information retrieval), and others. Many of the abovementioned systems already have integrated workflows for terminology management; therefore, the linking to the TaaS platform will offer a wider access to existing and shared terminology. The latter systems (starting from search engines) may also exploit term lists as seeds for acquiring Web data or to focus their search for data in domain-specific (or search query specific) directions. The Web-based API offers three main functions: lookup of terms in existing terminological collections, import of multilingual terminological collections, and export of collections from the TaaS platform. Terminological collections can be imported and exported using the TBX format. Additionally, for machine users that provide human users with the functionalities to search, create, delete and clean up terminology (like a terminology

management system), the TaaS platform offers advanced interfaces that operate similarly to the services offered for human users accessing the TaaS platform directly. In the TaaS project we study and evaluate the benefits of having access to multilingual terminology collections for two specific machine users. At the time of writing this demonstration paper, the TaaS platform's API interface was supported by the memoQ[5] CAT tool and the LetsMT[6] SMT platform (Vasiļjevs et al., 2012).

6. Conclusions

In this demonstration paper we have presented an innovative platform "Terminology as a Service" (TaaS) for acquiring raw terminological data, cleaning up, sharing, and reusing terminological data, based on cloud computing. The platform serves, among other things, the needs of specialised lexicography.

During the conference we will demonstrate the fully functional prototype of the platform. The live demonstration workflow will include extraction of terms from user-provided documents, as well as finding corresponding translation equivalents in terminology databases and in statistically aligned corpus data.

7. Acknowledgments

The work within the TaaS project has received funding from the European Union under grant agreement n° 296312.

We would like to thank the development team, our colleagues Andis Lagzdiņš and Pēteris Ņikiforovs, for their work on the TaaS platform.

8. References

Busa, R. (1961). Les travaux du Centre per l'automazione dell'analisis letteraria. *In Cahiers de Lexicologie*. Vol. 26. No 1.

Gornostay, T., Vopodiyanova, O., Vasiļjevs, A., & Schmitz, K.-D. (2013). Cloud-Based Terminology Services for Acquiring, Sharing and Reusing Multilingual Terminology for Human and Machine Users. *Proceedings of the conference TRALOGY II: Futures in Technologies for Translation. The quest for meaning: where are our weak points and what do we need?* Paris.

Blancafort, H., Daille, B., Gornostay, T., Heid, U., Méchoulam, C., & Sharoff, S. (2010). TTC: Terminology extraction, translation tools and comparable corpora. In *Proceedings, 14th EURALEX International Congress,* pp. 263-268.

Massion F. Управление терминологией: роскошь или необходимость?

[5] memoQ is available at: http://kilgray.com/products/memoq.

[6] LetsMT is accessible at: https://www.letsmt.eu.

Профессиональный перевод. Выпуск 12, 2007.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., & Gornostay, T. (2012). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)*, pp. 193-208. Madrid.

Pinnis, M. (Tilde), Ion, R., Ştefănescu, D., Su, F., Skadiņa, I. (Tilde), Vasiļjevs, A. (Tilde), & Babych, B. (2012). ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 91-96. Jeju: Association for Computational Linguistics.

Shaikevich, A. (1983). *Проблемы терминологической лексикографии = Problems of the Terminological Lexicography*. Moscow.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Volume 28, pp. 11-21.

Vasiljevs, A., Rirdance, S., & Liedskalnins, A. (2008). EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources* (ICGL 2008), pp. 213-220. Hong Kong

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: a cloud-based platform for do-it-yourself machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 43-48. Jeju: Association for Computational Linguistics.